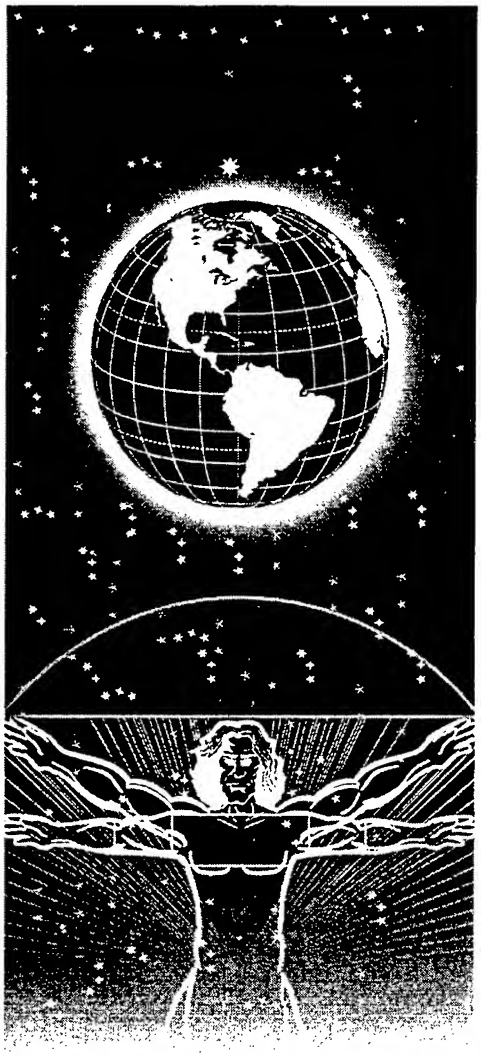
AL/HR-TP-1997-0005



UNITED STATES AIR FORCE RESEARCH LABORATORY

FACTOR STRUCTURE OF THE AIR FORCE OFFICER QUALIFYING TEST: ANALYSIS AND COMPARISON

Thomas R. Carretta
Aircrew Performance Branch
Aircrew Training Research Division

Malcolm J. Ree

Cognition and Performance Division

HUMAN RESOURCES DIRECTORATE
7909 Lindbergh Drive
Brooks AFB TX 78235-5352

June 1998

AIR FORCE MATERIEL COMMAND
AIR FORCE RESEARCH LABORATORY
HUMAN RESOURCES DIRECTORATE
AIRCREW TRAINING RESEARCH DIVISION
6001 South Power Road, Building 558
Mesa AZ 85206-0904

NOTICES

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

THOMAS R. CARRETTA
Project Scientist

DEE H. ANDREWS
Technical Director

LYNN A. CARROLL, Col, USAF
Chief, Warfighter Training Research Division

Please notify AFRL/HEOP, 2509 Kennedy Drive, Bldg 125, Brooks AFB, TX 78235-5118, if your address changes, or if you no longer want to receive our technical reports. You may write or call the STINFO Office at DSN 240-3877 or commercial (210) 536-3877.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 1998	Interim - August 1995 to October 1995

4. TITLE AND SUBTITLE

Factor Structure of the Air Force Officer Qualifying Test: Analysis and Comparison

5. FUNDING NUMBERS

PE - 62205F
PR - 1123
TA - B1
WU - 01

6. AUTHOR(S)

Thomas R. Carretta
Malcolm J. Ree

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division, Training Effectiveness Branch
7909 Lindbergh Drive
Brooks Air Force Base TX 78235-5352

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6001 South Power Road, Bldg 558
Mesa AZ 85206-0904

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL/HR-TP-1997-0005

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Dr Thomas R. Carretta, (210) 536-3956, DSN 240-3956
This paper was published previously in Military Psychology, 8, 29-42.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Air Force Officer Qualifying Test (AFOQT) is used to qualify men and women for commissions in the Air Force, classify them for pilot and navigator jobs, and award Reserve Officer Training Corps (ROTC) scholarships. Despite more than three decades of use, little published literature exists outside of Air Force technical reports that do not receive wide distribution. One of the most important details about a test battery is which factors it measures. To determine this, several factor models were tested with structural equations. Most of the models were hierarchical with general cognitive ability (g) as the highest factor. A model with hierarchical g and the five lower-order factors of verbal, math, spatial, aircrew, and perceptual speed fit the data best. The factor structure of the AFOQT was compared to the factor structure of the Armed Services Vocational Aptitude Battery (ASVAB), the enlistment qualification test battery. The AFOQT was found to contain a greater number of factors than the ASVAB. Given the confirmed AFOQT factor structure, four methods of increasing its validity were proposed and discussed. These methods were: increasing reliability of the tests, increasing the g-saturation, adding job knowledge tests, and adding additional valid factors.

14. SUBJECT TERMS

AFOQT; Air Force Officer Qualifying Test; Armed Services Vocational Aptitude Battery; ASVAB; Confirmatory factor analysis; Factor structure; Personnel measurement; Pilot selection; Pilot trainng; Navigator training; Training;

15. NUMBER OF PAGES
22

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION ABSTRACT
Unclassified	Unclassified	Unclassified	UL

CONTENTS

	Page
INTRODUCTION	1
METHOD	3
Participants	3
Measures	3
Procedures	5
RESULTS AND DISCUSSION	6
Increasing Reliability	10
Increasing *g*-Saturation	11
Adding Job Knowledge Tests	11
Adding Valid Factors	12
CONCLUSIONS	12
REFERENCES	13

FIGURE AND TABLES

Figure No.

1	The Factor Structure of the AFOQT and the ASVAB	9

Table No.

1	Composition of AFOQT Aptitude Composites	4
2	Correlation Matrix of the AFOQT Tests	7
3	Factor Loadings of the AFOQT Models	8
4	Factor Loadings of the AFOQT for Model 5	8

PREFACE

This effort was conducted under Work Unit 1123-B1-01, Pilot Selection and Classification Support, which is dedicated to research into the selection and classification of United States Air Force aircrew personnel. This paper was published previously as Carretta, T. R., & Ree, M. J. (1996), titled Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison, in *Military Psychology, 8,* 29-42.

FACTOR STRUCTURE OF THE AIR FORCE OFFICER QUALIFYING TEST: ANALYSIS AND COMPARISON

INTRODUCTION

The Air Force Officer Qualifying Test (AFOQT) is used for the commissioning and classification of United States Air Force officers through the Reserve Officer Training Corps (ROTC) and Officer Training School (OTS). The AFOQT is also used to qualify applicants who pass other physical and educational requirements for pilot and navigator training. It has been validated for more than three dozen other officer jobs.

For operational use, the Air Force aggregates the 16 tests into five distinct but overlapping composites (see Table 1). As their names suggest, the Pilot and Navigator-Technical composites are used to qualify applicants for pilot and navigator training while the Academic Aptitude, Verbal, and Quantitative composites are used to select applicants into ROTC and OTS commissioning programs.

The AFOQT has been validated for the selection of pilots and navigators (Arth, Steuck, Sorrentino, & Burke, 1990; Carretta, 1992; Carretta & Ree, 1994a; Koonce, 1982; Olea & Ree, 1994), and other officer jobs (Arth, 1986; Arth & Skinner, 1986; Finegold & Rogers, 1985). For pilot and navigator training, prediction criteria included passing/failing training, average flying performance rankings, composites of specific check ride performances, and day and night celestial navigational tasks.

For eight non-flying officer jobs, Arth and Skinner (1986) demonstrated that the Academic Aptitude composite, a highly *g*-saturated measure, was valid (average uncorrected r = .31) for predicting final technical training grades. Similarly, Finegold and Rogers (1985) demonstrated the validity of all the AFOQT composites for predicting technical training grades, course completion, and course rank for Air Weapons Controllers (r ranged from .19 to .39). Arth (1986) demonstrated the validity of all the composites for predicting final technical school grades in almost all of 37 officer technical training schools investigated.

Olea and Ree (1994) have shown that the major source of validity for the AFOQT comes from its measurement of *g*. They also have shown that the incremental validity of the AFOQT tests beyond *g* was about .02 (.462 versus .482 for the summed composite criterion) for navigators across five individual training criteria and a summed composite of the five training criteria. The incremental validity of the non-*g* portions of the AFOQT for predicting pilot training success was about .08 (.314 versus .398 for the summed composite criterion). Most of this increment was due to tests measuring specific job knowledge (Hunter, 1983) about aircraft,

flying, and flight instruments. For a historical review on the role of *g* in military pilot selection, see Ree and Carretta (1996).

Skinner and Ree (1987) conducted exploratory factor analyses of the AFOQT on a sample of 3,000 Air Force officer commissioning applicants and described a five-factor solution: verbal, math, spatial, aircrew interest/aptitude, and perceptual speed. They used a principal factors analysis with communalities in the principal diagonal and Kaiser-Harris Type II oblique rotation. The factors all correlated positively with an average correlation of .36 and a range of .22 to .50.

Earles and Ree (1991), noting an average test intercorrelation of .43 in the Skinner and Ree (1987) data (with a range of .17 to .77) and the positive manifold of the factor correlations, extracted a hierarchical estimate of *g* using several sets of lower-order factors. Like Skinner and Ree (1987), they did not confirm the fit of the model through confirmatory factor analysis. For an annotated bibliography of studies on the AFOQT, see Cowan and Sperl (1989).

The Air Force also uses the Armed Services Vocational Aptitude Battery (ASVAB) to assess applicants for enlisted jobs. The structure of the ASVAB includes verbal/math, speed, and technical knowledge lower-order factors with *g* in a hierarchical position (Ree & Carretta, 1994). In the ASVAB, *g* accounts for 64% of the total variance. The lower-order factors account for 16% of the total variance as follows: speed, 6%, verbal/math, 2%, and technical knowledge, 8%.

The ASVAB's reliability and validity has been established for numerous criteria including training success (Earles & Ree, 1992), job performance work samples (Ree, Earles, & Teachout, 1994; Teachout & Pellum, 1991), and supervisory ratings of job performance, job knowledge tests, and work samples (McHenry, Hough, Toquam, Hanson, & Ashworth , 1990). The major source of validity for the ASVAB comes from its measurement of *g*. Ree and Earles (1991) and Ree, Earles, and Teachout (1994) have shown that the non-*g* portions of ASVAB increment validity correlations by only a small amount, about .02, regardless of the criterion.

The non-*g* portions of ASVAB include measures of technical knowledge, which provide almost all of the incremental validity (see Hunter, 1983; Ree & Earles, 1991). In a meta-analytic path analysis of 14 disparate jobs with 3,264 participants, Hunter (1983) showed that *g* influenced job knowledge, and job knowledge influenced job performance. Included in the measures of job performance were work samples and supervisory ratings. This path model suggested that job knowledge validity is a consequence of *g*. In a meta-analysis of the validity of job knowledge tests, Dye, Reck, and McDaniel, (1993) found that the validity generalized across a broad variety of jobs but was moderated by job complexity. Greater validity was found for more complex jobs. They also found that the more similar the content of the job knowledge test to the knowledge required on the job, the more valid the job knowledge test was. Job knowledge tests in a test battery could be expected to increase its validity beyond the validity offered by *g*,

especially for complex jobs or if the job knowledge were closely related to the content of the training or the job performance.

Despite long, common use by the same organization, the relationship of the ASVAB and the AFOQT has been investigated only once. Sperl, Ree, and Steuck (1992) have shown that the relationship of all the ASVAB tests and the verbal and mathematical tests of the AFOQT was due to the common measurement of *g*. They were unable to investigate the other AFOQT tests because the available enlisted Air Force participants, on average, scored at or near chance levels on these tests.

The purpose of the present study was to establish and confirm a factor structure for the AFOQT and to compare it with the ASVAB. Because the AFOQT and ASVAB differ markedly in difficulty, no single sample could take both tests (Sperl, Ree, & Steuck, 1992). The current study was conducted by analyzing AFOQT data on a sample of applicants to officer commissioning programs, and to compare the AFOQT factors to ASVAB factors confirmed in a previous study of the ASVAB normative sample (Ree & Carretta, 1994). Implications for the validity of the AFOQT as a consequence of its factor structure were considered.

METHOD

Participants

The participants, taken from Skinner and Ree (1987), were a random sample of 3,000 applicants for Air Force commissions selected from 126,747 examinees tested between late 1981 and 1985. Eighty-four percent were men, 78% were White, and 91% had completed either high school or college.

Measures

The AFOQT consists of 16 tests and 5 composites (Table 1). The correlations among these tests in the sample appear in Table 2. Brief descriptions of the AFOQT tests grouped according to content are provided below.

Verbal tests. Verbal Analogies (VA) measures ability to reason and recognize relationships between words. Reading Comprehension (RC) assesses reading skill. Word Knowledge (WK) provides a measure of ability to understand written language through the use of synonyms.

Quantitative tests. Arithmetic Reasoning (AR) measures understanding of arithmetic relationships expressed as word problems. Data Interpretation (DI) measures the ability to extract data from graphs and charts. Math Knowledge (MK) consists of items requiring the use of mathematical terms, formulas, and relationships.

Table 1. Composition of AFOQT Aptitude Composites

Test	Abbr.	Composite				
		Verbal	Quantitative	Academic Aptitude	Pilot	Navigator-Technical
Verbal Analogies	VA	X		X	X	
Arithmetic Reasoning	AR		X	X		X
Reading Comp.	RC	X		X		
Data Interpretation	DI		X	X		X
Word Knowledge	WK	X		X		
Math Knowledge	MK		X	X		X
Mechanical Comp.	MC				X	X
Electrical Maze	EM				X	X
Scale Reading	SR				X	X
Instrument Comp.	IC				X	
Block Counting	BC				X	X
Table Reading	TR				X	X
Aviation Information	AI				X	
Rotated Blocks	RB					X
General Science	GS					X
Hidden Figures	HF					X

Spatial tests. Mechanical Comprehension (MC) assesses understanding of mechanical functions. Electrical Maze (EM) measures spatial ability based on choice of a path through a maze. Block Counting (BC) assesses spatial ability through analysis of three-dimensional representations of a set of blocks. Rotated Blocks (RB) measures spatial aptitude by requiring mental manipulation and rotation of objects. Hidden Figures (HF) measures spatial ability by requiring the discovery of simple figures embedded in complex drawings.

Aircrew tests. Instrument Comprehension (IC) assesses the ability to ascertain aircraft attitude from illustrations of flight instruments. Aviation Information (AI) measures knowledge of general aviation terminology and concepts. General Science (GS) assesses knowledge and understanding of scientific terms, concepts, principles, and instruments.

Perceptual speed tests. Scale Reading (SR) measures the ability to read dials and scales. Table Reading (TR) assess the ability to quickly and accurately extract information from tables.

Procedures

Several models were specified, estimated using maximum likelihood procedures, and fit to the data (Bentler, 1989, 1990). Model 1 was defined by the operational composites not including Academic Aptitude, which would have been redundant and would have created a linear dependency in the data. Model 2 sought to confirm the five-factor structure found by Skinner and Ree (1987). This model included a verbal factor (VA, RC, WK, GS), a math factor (AR, DI, MK, SR), a spatial factor (MC, EM, BC, RB, HF), an aircrew factor (MC, IC, AI, GS), and a perceptual speed factor (DI, SR, BC, TR).

Model 3 included only a single common factor, psychometric *g*. All the remaining models used a hierarchical *g* factor with residualized lower-order factors (Schmid & Leiman, 1957). Model 4 was *g* and the operational composites, while Model 5 was *g* and the Skinner and Ree (1987) lower-order factors. Model 6 was a simplification of Model 5 achieved by combining the spatial and perceptual speed factors. Model 7 combined verbal and math factors to yield a Vernon-like model (Vernon, 1969) with hierarchical *g* and residualized lower-order verbal-math (combined from Skinner & Ree, 1987), spatial-perceptual speed (combined from Skinner & Ree, 1987), and aviation technical knowledge factors.

Several fit statistics and the standardized residuals were evaluated for each model. Marsh, Balla, and McDonald (1988) have shown that the popular Bentler-Bonnett non-normed index may be susceptible to sample size effects. They recommend evaluation of the Tucker-Lewis incremental fit index (TLI). Bentler (1990) has developed the Comparative Fit Index (CFI) based on the TLI and has shown that it is less dependent on sample size and has a smaller sampling variance than the TLI. Additionally, the average absolute standardized residual (AASR) from the residual correlation matrix was computed. Several indexes were used to evaluate models for goodness-of-fit, including CFI, the model chi-square statistic, Root Mean Square Error of Approximation (RMSEA) (Browne & Cudeck, 1993), AASR, and parsimony (Ree & Carretta, 1994). The Bentler-Bonnet CFI (Bentler, 1989) was used because it provided an accurate estimate of fit with low sampling variance. This index is scaled between zero and one, with higher values indicating better fit. The RMSEA was computed to show deviation of fit per degree of freedom. Chi-squares and residuals were inspected for each model.

Only nested models could be compared directly. In nested models, the factors in the simpler model are a subset of the factors in the more complex model. Goodness-of-fit was compared for the two nested models with the highest CFI (and lowest RMSEA). This was done by subtracting the model chi-squares ($\chi_D^2 = \chi_1^2 - \chi_2^2$) and evaluating with degrees of freedom equal to the difference in the degrees of freedom for the two model chi-squares. A significant χ_D^2 indicates a better fit for the model with the greater number of parameters.

RESULTS AND DISCUSSION

The correlations used to estimate the structural models are shown in Table 2. They are all positive with the lowest correlation (.17) between EM and WK and the highest correlation (.77) between WK and RC.

All the models were estimated with no problems in parameter estimates. Fit statistics, number of parameters estimated, and degrees of freedom are shown in Table 3.

The *g*-only model and the two non-hierarchical models showed relatively poor fit to the data. Model 1, which embodied the four operational composites, showed a CFI of .808, an RMSEA of .136, and an AASR of .196, suggesting a poor fit. Model 2, a five-factor non-hierarchical model, likewise showed a poor fit with CFI of .741, RMSEA of .154, and AASR of .293.

The single-factor model (Model 3) positing only *g* showed a relatively poor fit with a CFI of .779, RMSEA of .139, and AASR of .054. Model 4, which had *g* and factors mimicking the operational composites, showed a CFI of .949, RMSEA of .078, and AASR of .028, while Model 5 produced a CFI of .957, RMSEA of .071, and AASR of .027. Models 6 and 7 had CFIs of .954 and .947, RMSEAs of .072 and .077, and AASRs of .029 and .027, respectively.

The two models with the highest CFIs (and lowest RMSEAs), nested Models 5 and 6, were compared in a chi-square analysis. Results showed that there was a significantly better fit for Model 5. In Model 5, the proportion of total variance accounted for by *g* was 41%. The residualized lower-order factors accounted for 20% of the total variance as follows: verbal, 7%, math, 2%, spatial, 3%, aircrew, 5%, and perceptual speed, 3%. The factor loadings for Model 5 are shown in Table 4.

Clearly, most of the fit comes from the *g* factor. In the *g*-only model, *g* accounted for 51% of the total variance. When the lower-order common factors were added, the fit improved as expected. In each of the hierarchical models, *g* accounted for more of the total variance (41%) than did the sum of the lower-order factors. This was expected (Jensen, 1980). However, *g* accounted for a smaller proportion of the factor and total variance in the AFOQT (41% of the total variance) than in the ASVAB (64% of the total variance) (Ree & Carretta, 1994).

In comparison to the ASVAB (Ree & Carretta, 1994), the AFOQT measured a greater number of factors (Figure 1). The AFOQT provided separate factors for verbal and math instead of a single verbal-math factor. This may be because the AFOQT has four verbal and four math tests but ASVAB has only two of each. It should be remembered that a properly determined factor has at least three indicators.

Table 2. Correlation Matrix of the AFOQT Tests.

Test	VA	AR	RC	DI	WK	MK	MC	EM	SR	IC	BC	TR	AI	RB	GS	HF
VA	100															
AR	58	100														
RC	73	58	100													
DI	53	67	55	100												
WK	68	46	77	46	100											
MK	55	71	51	60	40	100										
MC	48	51	46	46	40	48	100									
EM	27	37	23	38	17	40	44	100								
SR	48	66	45	62	37	60	48	51	100							
IC	34	41	33	43	28	39	49	44	49	100						
BC	45	53	40	51	32	49	50	47	61	49	100					
TR	34	44	35	47	27	44	30	31	56	34	51	100				
AI	30	31	34	34	32	25	50	29	33	56	31	21	100			
RB	43	47	35	42	29	49	54	42	49	46	55	34	34	100		
GS	51	49	55	44	51	52	57	34	41	41	37	25	46	40	100	
HF	40	40	36	39	31	40	39	34	37	36	45	36	27	42	34	100

Note. See Table 1 for definitions of test name abbreviations

Table 3. Fit Statistics for AFOQT Models.

Model	df	# of Estimated Parameters	χ^2	RMSEA	CFI	AASR
1	95	41	5250	.136	.808	.196
2	99	37	7052	.154	.741	.293
3	104	32	6027	.139	.779	.054
4	79	57	1448	.078	.949	.028
5	83	53	1250	.071	.957	.027
6	84	52	1313	.072	.954	.029
7	84	52	1509	.077	.947	.027

N = 3,000

Table 4. Factor Loadings of the AFOQT for Model 5.

	Factor					
Test	*g*	Verbal	Math	Spatial	Aircrew	Perceptual Speed
VA	.683	.460				
AR	.781		.520			
RC	.650	.617				
DI	.731		.189			.131
WK	.532	.690				
MK	.760		.224			
MC	.667			.224	.265	
EM	.527			.273		
SR	.758		.123			.305
IC	.588				.384	
BC	.672			.328		.299
TR	.544					.485
AI	.437				.766	
RB	.634			.383		
GS	.618	.228			.238	
HF	.562			.151		
% Total Variance						
	41.1	7.0	2.3	2.5	5.4	2.7

Note. See Table 1 for definitions of test name abbreviations. The loadings have been residualized so that the effects of the higher-order factor, *g*, have been removed from the lower-order factors. Approximately 39% of the variance was due to test-specific uniqueness and unreliability.

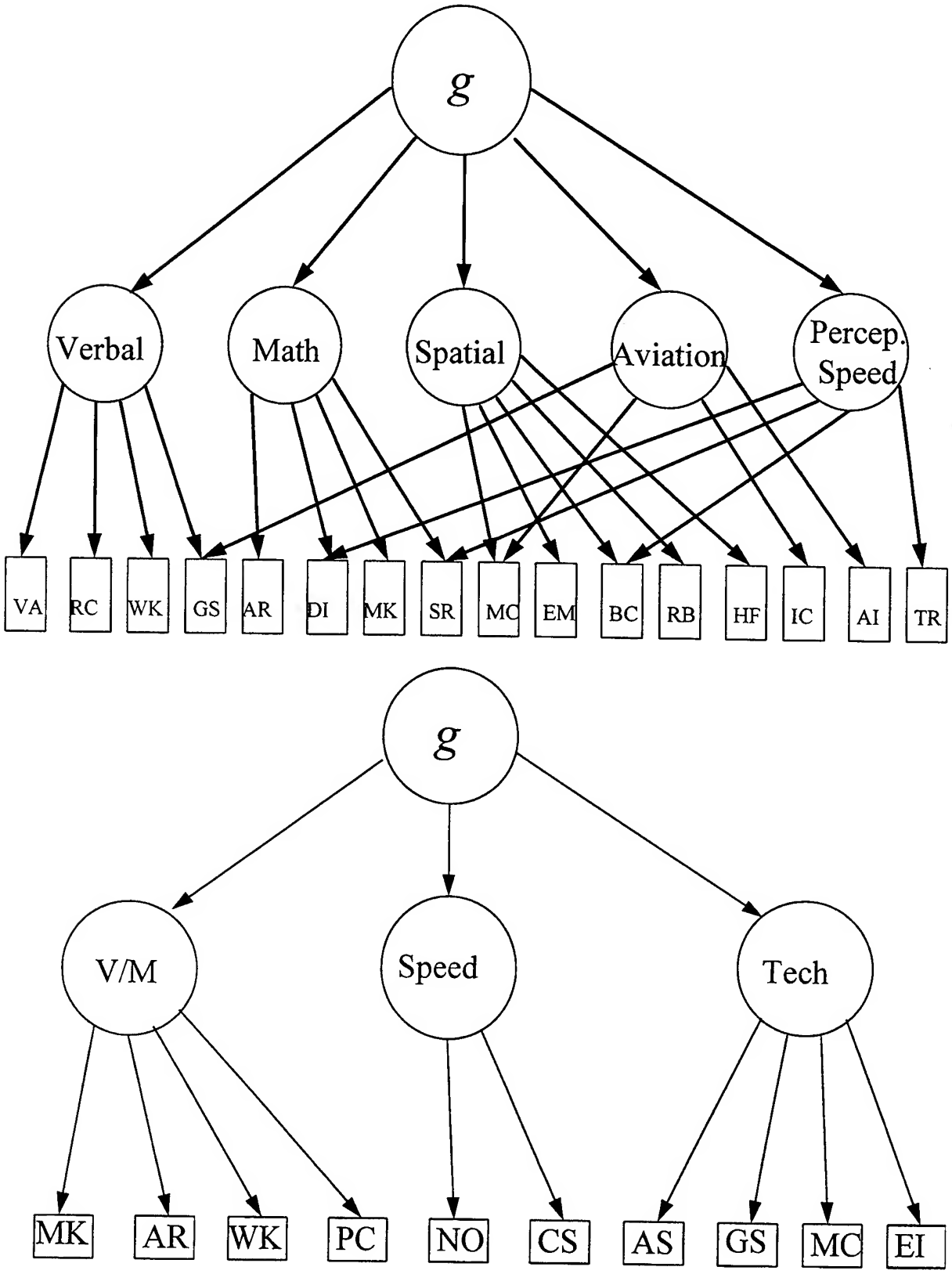


Figure 1. The Factor Structure of the AFOQT and the ASVAB

Note. ASVAB has 10 tests including General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI).

In the AFOQT there were spatial and aircrew factors not found in ASVAB. There is nothing comparable to the spatial factor of AFOQT in the ASVAB. However, the aircrew factor is similar to the technical knowledge factor in the ASVAB (Ree & Carretta, 1994) because both measure, in large part, acquired job knowledge not necessarily found in the common core of the usual educational curriculum.

While job knowledge tests are useful for prediction (Dye et al., 1993), if they are too specific, their predictiveness may be perishable as technology changes. For example, questions found on earlier versions of the ASVAB could not be used today because their content reflected vanished technologies such as transformers in audio output circuits and "points" in automobile ignition systems. In contrast, the AFOQT has generally avoided such specificity by including only items covering broad areas of knowledge about aircraft. In the long run, general job knowledge test questions may prove preferable to highly specific job knowledge test questions.

Another issue concerning job knowledge test validity is the potential for trainees without the knowledge to catch up to those with it, if training includes it and is long enough. Carretta and Ree (1994a) have pointed out that the job knowledge tests of the AFOQT, AI and IC, are more valid for early phases of pilot training than for later phases. ASVAB studies across jobs tend not to show this effect. For example, Ree and Earles (1991) found an average increment for the non-*g* portions of the ASVAB of about .02 for predicting early training success. McHenry et al. (1990) and Ree, Earles, and Teachout (1994) found about the same increment for predicting measures of job performance collected after several years on the job. It is likely that the specialized job knowledge tests of the ASVAB contain information that is highly related to only a few of the many enlisted jobs, and that the knowledge gained in technical training and on the job is more important to job performance than job knowledge at time of testing. Further, the longer term effects of predicting with job knowledge tests are unknown. It is not known how long during a military career job knowledge tests would be valid predictors of performance.

There are four broad changes that might improve the validity of AFOQT. The first would be to increase the reliability of the tests. The second would be to increase the *g*-saturation of the battery. The third would be to increase the proportion of variance attributable to job knowledge. The fourth would be to incorporate tests that increase the range of valid factors measured.

Increasing Reliability

Increasing the reliability offers little hope of making dramatic changes in the validity of the AFOQT. The reliability of the AFOQT tests is generally high with a median coefficient alpha of .80 and a range of .69 to .92 (Skinner & Ree, 1987). Only highly reliable composites are operationally used for personnel selection and classification (see Table 1).

Increasing g-Saturation

Ree and Earles (1994) reported that the correlation between the *g*-saturation of a test and its average validity was .96 (see also Jensen, 1980). Therefore, if *g*-saturation is increased, an increase in validity is predicted. An increase in the *g*-saturation of the AFOQT could be accomplished by removing tests low in *g* and replacing them with tests high in *g*. There are at least three ways in which this could be accomplished. The first, and perhaps least desirable way, would be to include content that is dependent on discretionary activities in which all do not participate. These discretionary activities might be hunting, fishing, cosmetics, or various sports. Other sources of content could be current events, but this has proven to be perishable in former versions of the ASVAB. For example, general knowledge tests have been used on both the AFOQT and the ASVAB and have been abandoned because the knowledge proved not to be as general as believed and certain group mean differences were unacceptable.

The second way to increase *g*-saturation would be to use tests that reflect content common to all educational curricula. These are content areas that all examinees will have studied at least to some extent. Examples include verbal reasoning, word knowledge, reading skill, mathematical reasoning, and mathematical knowledge. Tests like these are already in the AFOQT, but additional similar tests could be added.

A third way to increase *g*-saturation would be to add tests that do not depend on learned content, such as Raven's Matrices (1966) or elementary cognitive tasks. Kranzler and Jensen (1991), Kyllonen (1993), and Kyllonen and Christal (1990) have demonstrated that elementary cognitive tasks measure *g* as well as paper-and-pencil tests and can be developed with no learned content. Further, analyses of batteries of elementary cognitive tasks show that *g* accounts for more of the variance among the intercorrelations than in paper-and-pencil multiple aptitude batteries.

Adding Job Knowledge Tests

Adding job knowledge tests is the third method to increase the validity of the AFOQT. Currently, only AI and IC can reasonably be considered as job knowledge tests. Olea and Ree (1994) examined the validity and incremental validity of measures of *g*, specific abilities (e.g., verbal, spatial), and specific knowledge (i.e., pilot job knowledge) from the AFOQT for predicting several pilot criteria including training performance grades and flying-work samples. They observed that most of the predictive utility came from *g* and that *specific abilities* added little incremental validity beyond *g*. Most of the incremental validity to *g* in the AFOQT came from the pilot job knowledge tests (AI and IC). These tests would not be considered job knowledge tests for other occupations in which officers serve. It may be suggested that job knowledge tests that were not job specific, but job-family specific, might be useful. No research specifically in this area has been reported.

The strategy of including job knowledge tests for each Air Force officer job is problematic. There are numerous technical training courses in which officers enroll. To add a test for each training course to the AFOQT is impractical. A solution might be to develop a battery that is adaptive among tests, administering only those relevant to a particular course. This brings, among others, problems for producing equivalent forms of the battery. Regardless of how many individuals are administered a particular test, it must be equated, validated, and examined for bias. Further, keeping job knowledge tests current is expensive and time consuming. Dye et al. (1993) showed that job knowledge tests achieve maximum validity when the content is most similar to job content. This potentially requires that job knowledge test content be updated frequently. An additional complication is possible adverse impact, because all sex and ethnic groups are not expected to have acquired the same level of job knowledge prior to application.

Adding Valid Factors

Increasing the number of *valid* factors measured by the AFOQT would be the fourth method to improve its validity. Numerous studies have shown *g* to be the ubiquitous predictor of job and training performance (Hunter & Hunter, 1984; McHenry et al., 1990; Olea & Ree, 1994; Ree & Earles, 1991, 1992, 1993; Ree, Earles, & Teachout, 1994; Schmidt, Hunter, & Pearlman, 1981). Hunter and Hunter have shown that for certain jobs, psychomotor skills are predictive and Carretta and Ree (1994b) have shown that psychomotor skills are predictive of pilot training success. Likewise, Tett, Jackson, and Rothstein (1991) and Mount and Barrick (1991) have demonstrated the generalizable validity of personality measures. McHenry et al. (1990) have shown the incremental predictive validity of temperament/personality (shown in parentheses are validity for *g* versus validity for *g* and temperament/personality) for specific criteria such as effort and leadership (.31 to .42), personal discipline (.16 to .35), and physical fitness and military bearing (.20 to .41). Combining measures of psychomotor skills and temperament/personality with all or part of the AFOQT could improve validity.

CONCLUSIONS

The AFOQT is comprised of five lower-order factors: verbal, math, spatial, aircrew, and perceptual speed. These accounted for 20% of the total variance, and *g* in hierarchical position accounted for 41% of the total variance. Compared with the ASVAB, the AFOQT was less *g*-saturated but had more common factors and had a greater proportion of its variance associated with common factors. Four methods of increasing the validity of the AFOQT were proposed and discussed. These methods were: increasing reliability of the tests, increasing *g*-saturation, adding job knowledge tests, and adding additional valid factors. Future forms of the AFOQT could be built around a core of *g*-saturated tests, with a block of job knowledge tests, psychomotor tests, and measures of temperament/personality (Carretta & Ree, 1994b).

REFERENCES

Arth, T. O. (1986). *Validation of the AFOQT for non-rated officers* (AFHRL-TP-85-50). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Arth, T. O. , & Skinner, M. J. (1986, November). *Aptitude selectors for Air Force officer non-aircrew jobs.* Paper presented at the annual meeting of the Military Testing Association, Mystic, CN.

Arth, T. O., Steuck, K. W., Sorrentino, C. T., & Burke, E. F. (1990). *Air Force Officer Qualifying Test (AFOQT): Predictors of undergraduate pilot training and undergraduate navigator training success* (AFHRL-TP-89-52). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Lang (Eds.) *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Carretta, T. R. (1992). Understanding the relations between selection factors and pilot training performance: Does the criterion make a difference? *The International Journal of Aviation Psychology, 2,* 95-105.

Carretta, T. R., & Ree, M. J. (1994a). Air Force Officer Qualifying Test validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379-387.

Carretta, T. R., & Ree, M. J. (1994b). Pilot candidate selection method (PCSM): Sources of validity. *The International Journal of Aviation Psychology, 4,* 103-117.

Cowan, D. K., & Sperl, T. C. (1989). *Selection and classification of United States military officers: A fifty-year bibliography (1937-1986)* (AFHRL-TP-88-45). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Dye, D. A., Reck, M., & McDaniel, M. (1993). The validity of job knowledge measures. *International Journal of Selection and Measurement, 1,* 153-162.

Earles, J. A., & Ree, M. J. (1991). *Air Force Officer Qualifying Test (AFOQT): Estimating the general ability component* (AL-TP-1991-0039). Brooks AFB, TX: Manpower and Personnel Research Division, Armstrong Laboratory, Human Resources Directorate.

Earles, J. A., & Ree, M. J. (1992). The predictive validity of ASVAB for training grades. *Educational and Psychological Measurement, 52,* 721-725.

Finegold, L., & Rogers, D. (1985). *Relationship between Air Force Qualifying Test scores and success in air weapons controller training* (AFHRL-TR-85-13). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland, (Eds.), *Performance Measurement and Theory* (pp. 257-266). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Jensen, A. R. (1980). Bias in mental testing. New York: The Free Press.

Koonce, J. M. (1982). Validation of a proposed pilot-trainee selection system. *Aviation, Space, and Environmental Medicine, 53,* 1166-1169.

Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric *g*: Unitary process or a number of independent processes? *Intelligence, 15,* 397-422.

Kyllonen, P. C. (1993). Aptitude testing inspired by information processing: A test of the four-sources model. *The Journal of General Psychology, 120,* 375-405.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!, *Intelligence, 14,* 389-433.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391-410.

McHenry, J. J., Hough, L. M, Toquam, J. L., Hanson, M., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

Mount, M. R., & Barrick, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Olea, M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-849.

Raven, J. C. (1966). *Advanced progressive matrices.* New York: Psychological Corporation.

Ree, M. J., & Carretta, T. R. (1994). Factor analysis of ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement, 54,* 459-463.

Ree, M. J., & Carretta, T. R. (1996). Central role of *g* in military pilot selection. *The International Journal of Aviation Psychology, 6,* 111-123.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 321-332.

Ree, M. J., & Earles, J. A. (1992) Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86-89.

Ree, M. J., & Earles, J. A. (1993). *g* is to psychology what carbon is to chemistry. A reply to Sternberg and Wagner, and to McClelland and Calfee. *Current Directions in Psychological Science, 2,* 11-12..

Ree, M. J., & Earles, J. A. (1994). The ubiquitous predictiveness of *g*. In C. Walker, M. Rumsey, & J. Harris (Eds.), *Personnel selection and classification* (pp. 127-135). Hillsdale, NJ: Earlbaum.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518-524.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66,* 166-185.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O* (AFHRL-TR-86-68). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sperl, T. C., Ree M. J., & Steuck K. R. (1992) Armed Services Vocational Aptitude Battery and Air Force Officer Qualification Test: Analyses of common attributes. *Military Psychology, 4,* 175-186.

Teachout, M. S., & Pellum, M. W. (1991). *Air Force research to link standards for enlistment to on- the-job performance* (AFHRL-TR-90-90). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703-742.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Vernon, P. E. (1969). *Intelligence and cultural environment.* London: Methuen.